DOCUMENT RESUME

ED 178 575 TM 009 730

AUTHOR Kukuk, Cris R.; And Others
TITLE An Analysis with Illustrations of Suppressor Variables in Social Science Research.
PUB DATE Apr 79
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, CA, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Correlation; Criteria; *Multiple Regression Analysis; *Predictor Variables; Research Design; Social Science Research; *Suppressor Variables

ABSTRACT

The purposes of this paper are to provide an explanation of the way in which suppressor variables operate in multiple regression, and to provide illustrations of their effects with actual social science data. Examples are reviewed in which one predictor accounts for more of the variance in the dependent variable when another predictor is included, and in which the last variable to enter in a stepwise regression equation accounts for more variance than the next-to-last variable. Examples include the following: (1) in predicting the number of doctors in an area from the number of hospital beds, the money spent on hospital facilities was a suppressor variable; (2) in predicting rural or urban preferences for location of the future practice of pharmacy students by its relationship with sex, age was a suppressor variable; (3) in predicting reading scores by a personality inventory, personal integration scores were a suppressor variable for the estheticism score; and (4) in the prediction of the larceny rate in certain census tracts, the robbery rate was a suppressor variable for a socioeconomic indicator. Analysis of variance may be used to clarify suppression. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178575

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AN ANALYSIS WITH ILLUSTRATIONS OF SUPPRESSOR VARIABLES IN SOCIAL SCIENCE RESEARCH

Cris R. Kukuk
Daniel U. Levine
Jeanie Keeny Meyer

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Chris Kukuk
Daniel U. Levine
Jeanie Keeny Meyer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association
April 8-12, 1979 San Francisco

TM 009 730

Although suppressor variables have been mentioned in social science literature for nearly forty years,[1] relatively little has been written about them and few actual examples have been described to help provide understanding of their importance and implications when they appear in multivariate research. The purposes of this paper are to provide a clear explanation of the way in which suppressor variables operate in multiple regression and to illustrate their effects with actual social science data.

Suppressor variables were defined by Horst as variables which improve the prediction of a criterion through the addition of a variable which is uncorrelated with the criterion but is related to another predictor or set of predictors.[2] More recently, Conger distinguished between three types of suppressors - "traditional," "negative," and "reciprocal" - and in so doing revised the definition of suppression to read as follows: "A suppressor variable is defined to be a variable which increases the predictive validity of another variable (or set of variables) by its inclusion in a regression equation. This variable is a suppressor only for those variables whose regression weights are increased."[3]

Most discussions of suppressor variables have been theoretical, but Rosenberg has provided some examples from actual data to help researchers understand how they may be important in practice.[4] "With a suppressor variable," Rosenberg points

---

[1] P. Horst, "The Prediction of Personal Adjustment," Social Science Research Council Bulletin, No. 48, New York, New York, 1941, cited in Anthony J. Conger, "A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation," Educational and Psychological Measurement, 34 (1974), 35-46.

[2] Ibid.

[3] Anthony J. Conger, op. cit., p. 37.

[4] Morris Rosenberg, "The Logical Status of Suppressor Variables," Public Opinion Quarterly, (1973), 359-372.

out, "one starts with a null correlation between two variables; when one controls on an antecedent variable, a relationship emerges in each of the test factor categories . . . this first-order relationship is real, it is the original null relationship that is spurious."[5] He illustrates this phenomenon with several examples such as one from Lipsitz' study of authoritarianism,[6] in which education was positively related to social class but negatively related to authoritarianism; the relationship between social class and authoritarianism did not emerge until their relationships with education were statistically controlled.

The necessity to control for intervening variables which may confound the relationship between variables of particular interest in a given study is, of course, widely recognized, but frequently researchers have little understanding of how suppressor variables may function in such a situation, or how to identify their presence and assess their importance. This lack of awareness exists partly because few sources exist which provide a theoretical understanding of suppressors using examples from actual data to show precisely how they function in the analysis of a data set, particularly when one uses the popular multiple regression approaches now available in packaged computer programs. This paper illustrates and explains the functioning and assessment of suppressors using actual data bearing on two manifestations of suppression in SPSS stepwise multiple regression analysis:

---

[5] Ibid., p. 363.

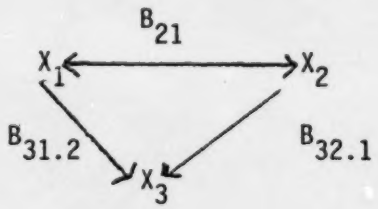
[6] L. Lipsitz, "Working Class Authoritarianism; a Re-evaluation," American Sociological Review, 30 (1965), 103-109.

1. A situation in which the relationship between X_3 and X_1 or X_2 is not significant (or is only marginally significant) but becomes significant (or more significant) when X_3 is predicted by both X_1 and X_2. In this situation X_1 is a suppressor variable for X_2, and X_2 is a suppressor variable for X_1.

2. A situation in stepwise multiple regression in which the addition of a (suppressor) variable explains more variance than that explained by a previously added variable. This situation can be considered an extension of the first situation but it is useful to analyze it separately.

ADDITION OF A PREDICTOR VARIABLE MAKES A PREVIOUS PREDICTOR SIGNIFICANT OR MORE SIGNIFICANT

Let us define the subscripts of a correlation or regression coefficient so that DI.C means the first symbol designates the dependent variable, the second symbol designates the independent variable, and any symbols following the period stand for control variables. Thus for variables X_1, X_2, and X_3, $r_{32.1}$ is the partial correlation coefficient between X_3 and X_2 controlling for X_1 with X_3 the dependent variable in a causal relationship portrayed as follows (the B's are standardized regression coefficients):



The relationship between X_1 and X_3 summarized by the correlation coefficient r_{31} decomposes into the following:

$$r_{31} = B_{31.2} + B_{21}B_{32.1}$$

The $B_{31.2}$ is the direct relationship between X_3 and X_1 controlling for X_2 while the product $B_2B_{32.1}$ is the indirect effect of X_1 on X_3 through the intervening variable X_2. Obviously, the absolute value of r_{31} is larger if the components are both of the same sign than if they are mixed and one is subtracted rather than added to the other. It is this latter situation which denotes a suppressor effect.

Let us consider a concrete example. We wish to predict the number of physicians (P) as a function of number of hospital beds (B) and money spent on hospital facilities (M) in forty-eight counties in the midwest.[7] All variables are standardized as rates per 1,000 population, and the regression coefficients, partial correlation coefficients, and zero-order correlations are as follows with P the dependent variable:[8]

Predictor Variable M

$$r_{PM} = B_{PM} = .06918,\ F = .22,\ P = ns$$

$$\left.\begin{aligned} r_{PM.B} &= .50294 \\ B_{PM.B} &= .77472 \end{aligned}\right\} F = 15.24,\ P < .01$$

Predictor Variable B

$$r_{PB} = B_{PB} = -.29257,\ F = 4.31,\ P < .05$$

$$\left.\begin{aligned} r_{PB.M} &= -.56001 \\ B_{PB.M} &= -.89994 \end{aligned}\right\} F = 20.56,\ P < .01$$

Our expectation is that spending money on hospital facilities (M) as well as having more hospital beds available (B) tends to attract physicians (P) to a county but that part of the effect of M on P is through its influence in increasing the number of hospital beds. Thus M and B both should have a positive correlation

---

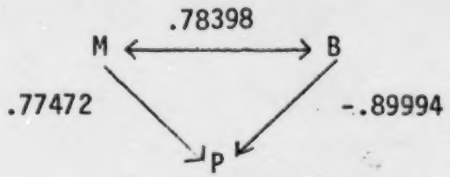
[7] Data for this example are taken from Bruce B. Morgan, Kenneth R. Mares, and Chris R. Kukuk, "Regression Analysis of Physician Distribution to Identify Areas of Need: Some Preliminary Findings." Paper presented at the Third Annual Data Use Conference of the National Center for Health Statistics, Phoenix, August, 1978.

[8] Coefficients in this paper are reported to five decimal places in order to obtain high accuracy in the computational demonstrations.

with P in accordance with the model shown above.

However, the zero-order correlation coefficients do not conform to expectations inasmuch as M is uncorrelated with P. The model shows why this is so.

The correlation between M and B is .78398. Since this is a bivariate correlation, B_{MP} also is .78398. Substituting M for X_1, B for X_2, and P for X_3 in our causal model, the model and regression coefficients are as follows:



The zero-order relationship between beds and physicians is statistically significant and negative (r_{PB} = .29257) but that relationship controlling for money is even more significant ($r_{PB.M}$ = .56001). The zero-order relationship between money and physicians is not significant (r_{PM} = .06918) but this relationship becomes significant ($r_{PM.B}$ = .50294) when the number of beds is controlled. The non-significant zero-order correlation between money and physicians is decomposed as follows:

$$r_{PM} = B_{PM.B} + B_{BM}B_{PB.m}$$
$$.06918 = .77472 + .78398\ (-.89994)$$
$$= .77472 - .70553$$
$$= .06918$$

These patterns develop because controlling for money spent on hospital facilities, there is a negative relationship between beds and physicians in this particular data set. This happens because money for facilities has been spent in counties that have a low number of physicians and an excess number of beds to begin with. Similarly, the supply of hospital beds has a small zero-order

correlation with physicians but the relationship becomes much stronger after controlling for money for facilities, because the money has been expended in counties that had relatively few physicians. The money spent on facilities does have a direct effect in attracting physicians, and an indirect effect through provision of beds in counties that had few physicians. The zero-order correlations are misleading because they do not take into account the negative relationship between beds and physicians which suppresses the relationships between the predictors and the dependent variable. Suppression in a three-variable model with two predictors will take place anytime the relationships between the predictors and the dependent variable differ in sign.

ADDITION OF A PREDICTOR ADDS MORE EXPLAINED VARIANCE TO A STEPWISE REGRESSION EQUATION THAN THE VARIANCE EXPLAINED BY THE PREVIOUS PREDICTOR.

Stepwise multiple regression is a process of selecting predictor variables successively based on their ability to explain variation in a dependent variable. Given a pool of potential predictors, the one which has the highest squared correlation with the dependent variable is selected first. Of the remaining pool, the variable selected next is the one with the highest squared partial correlation with the dependent variable when the first variable is controlled, and so on. This should yield a decreasing amount of variance added to the R^2 with each successive predictor.

Occasionally, however, the second variable adds more variance than did the first, or the last predictor in a set of three or more adds more variance than the next-to-last predictor. When this happens, it is either the result of a suppressor effect or of blow-up in the regression program. Since two predictor variables involved in a suppressor relationship mutually supress (i.e. spuriously reduce) each others' zero-order correlation with the dependent variable, the variable selected first may explain less variance than the one added second whose

explanatory power is assessed by the square of its partial correlation with the dependent variable, controlling for the first predictor.

An example is provided in a data set dealing with future practice location preferences of 56 university pharmacy students.[9] The study sought to identify the background factors which might predispose graduating pharmacy students to practice in rural locations. Students in the study were asked to designate their locational preferences and were classified as desiring or not desiring to locate in a rural area. Two of the variables which were included as predictor variables were sex and age. Data on the relationships between sex (S) and age (A; $r_{SA} = .32615$) and the dependent variable (R) were as follows:

Predictor Variable S

$$r_{RS} = B_{RS} = .25875$$

$$\left.\begin{aligned} r_{RS.A} &= .36765 \\ B_{RS.A} &= .37747 \end{aligned}\right\} F = 8.13,\ p < .01$$

Predictor Variable A

$$r_{RA} = B_{RA} = -.24088$$

$$\left.\begin{aligned} r_{RA.S} &= -.35622 \\ B_{RA.S} &= -.36399 \end{aligned}\right\} F - 7.56,\ p < .01$$

Since S had the higher correlation (.2578752) with R, it entered first, explaining 7 percent of the variance. The percent of variance added by A is the square of the partial correlation (-.35622), in this case 12 percent. Thus the second variable explained more of the variance than did the one selected first.

---

[9]The data are reported in Daniel U. Levine, Kenneth P. Mares, and Jane C. Pickett, "1978 Evaluation of the UMKC School of Pharmacy Externship Program." Kansas City, Missouri: Western Missouri Area Health Education Center, 1978.

Decomposition of the spuriously-low correlation coefficient of S with R is as follows:

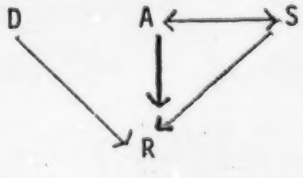
$$r_{RS} = B_{RS.A} + B_{SA}B_{RA.S}$$

$$.258752 = .37747 + .32615\ (-.36399)$$

$$.258752 = 37747 - .11872$$

$$= .25875$$

Adding another significant predictor to this data set throws further light on the way suppression can operate in a multivariate analysis. In this case we added scores on an attitude scale dealing with the distribution of drugs (D) to determine whether we could improve the prediction of rural location preference and whether the suppressor relationship between age and sex would be substantially affected. Attitudes on the distribution of drugs correlated at .29 with rural preference, so we knew this variable would enter first in a three-predictor equation. Having noted that D was virtually uncorrelated with age and sex (the respective zero-order correlations were .00 and .05), we expected that the following model would portray the relationships in the equation:



As expected, the model provided a good fit for the data, and the suppression relationship between the three predictor variables and the dependent variables in the stepwise regression equation were as follows

Dependent Variable R

Predictor Variable By Order of Entry	Zero-order r	Cumulative R^2	Increase in R^2	B	F
D	-.29	.08	.08	-.27187	5.076
S	.26	.14	.06	.36382	8.124
A	-.24	.26	.12	-.36062	7.998

The addition of another predictor thus did not change the pattern that age entered after sex but explained more variance. Betas between age and sex on the one hand and rural preference on the other are slightly changed due to the slight, non-significant relationship between D and rural preference, but decomposition according to our age, sex, and rural preference model continues to be highly accurate:

$$r_{RA} = B_{RA.S} + B_{AS}B_{RS.A}$$
$$-.24088 = -.36062 + .32615\ (.36382)$$
$$= -.36062 + .11866$$
$$= -.24196$$

It also is instructive to see how the partial correlations in this stepwise analysis are affected by the suppression relationship. As noted above, D entered the equation first, and at this point the partial correlations between S and R and A and R controlling for D were .25657 and -.25267, respectively. Since the partial correlation for S was slightly higher than for A, S entered the equation next, but the next set of partial correlations for stepwise inclusion showed that the partial correlation of A with R had increased to -.368. This increase was due to suppression, which explains why A entered the equation third but added more to the R^2 than did S.

ANALYSIS OF COVARIANCE

Since one of the predictor variables in the previous example dealing with age, sex and rural location preference of pharmacy students is interval and the other is nominal, the relationship between these predictors and the dependent can be examined using analysis of covariance. Analysis of covariance can further help in this case to show how a suppressor variable can increase the explanatory power of the predictor it suppresses. Examination of the substantive relationships involved in the covariance pattern illustrates how age and sex are related to choice

of a rural location in such a way as to cause suppression in a stepwise regression analysis.

The basic pattern is as follows: older students are less likely to prefer rural locations than younger students (rural preference was coded as 1, non-rural preference as zero; r = -.24), and male students are more likely to prefer rural locations than female students (male = 1, female = 0, r = .25). Moreover, the age-rural preference relationship holds within each of the sexes - the correlation between age and rural preference is -.37 for males and -.40 for females. As noted above, the correlation between age and rural preference for the total group is -.24. Thus age explains seven percent of the variance in rural preference.

Age and sex together, however, explain eighteen percent of the variance in rural preference, with sex entering the stepwise regression equation first because it has a higher zero-order correlation with rural preference than does age (see previous section). As shown in Figure 1, taking both age and sex into account in plotting the relationships among the variables provides more information than when age is plotted simply against rural preference.[10] The top line (on the left side) in the figure shows the slope of the regression line between age and rural preference for males, the bottom line shows the slope for females, and the middle line is for the total group. The top and bottom lines thus represent the age-rural preference relationship in an analysis of covariance controlling for sex. The middle line obviously provides less predictive power than the parallel top and bottom lines, which depart much more from a perfectly horizontal line that would represent no relationship between age and rural preference.

---

[10]The interaction term age X sex does not add significantly to the R^2.

SUPPRESSORS IN A LARGER SET OF PREDICTOR VARIABLES

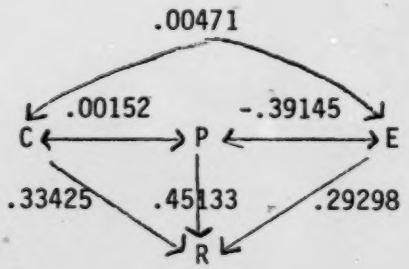
The next example illustrates the fact that one or more variables may enter in a stepwise regression before a suppressed variable is selected. A researcher[11] regressed reading percentile scores (R) on components of the Omnibus Personality Inventory (OPI). The first 3 variables entered were personal integration (P), complexity (C) and estheticism (E), in that order. The zero-order path coefficients were as follows:

$$r_{RP} = .38$$

$$r_{RC} = .34$$

$$r_{RE} = .12$$

The zero-order correlation .12 between estheticism and reading scores ranked seventh out of the thirteen components in magnitude. It was therefore somewhat surprising that it entered third in the regression. This zero-order relationship explains only 1.4 percent of the variance but when estheticism entered third in the stepwise regression run, the addition to explained variance was 7.3 percent. Clearly a suppressor relationship was operating, as in the following path model:



The suppressor relationship is between personal integration and estheticism but complexity enters before personal integration or estheticism because it has a higher partial correlation with the dependent variable. Complexity is essentially uncorrelated with the other two predictor variables:

---

[11] Deanna Martin, "Predicting Reading Achievement in College Students." Unpublished Ph.D. Dissertation, University of Missouri-Kansas City School of Education, 1978.

$$r_{CP} = .002$$

$$r_{CE} = .005$$

The zero-order correlation between estheticism and reading scores decomposes as follows:

$$r_{RE} = B_{RE.PC} + B_{PE}B_{RP.EC} + B_{CE}B_{RC.PE}$$

$$.11788 = .29398 + (-.39145)\ (.45133) + (.00471)(.33425)$$

$$= .29398 - .17667 + .00016$$

$$= .29414 - .17667$$

$$= .11747^{12}$$

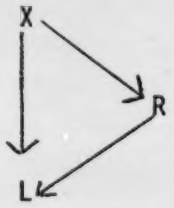
$$= .00041$$

---

[12] Note only paths going through one or less double headed arrows are theoretically possible; hence the unaccounted part of the correlation .00041 (.11788 - .11747) would be accounted for by more complex but trivial paths. The path through complexity (.00016) is also trivial.

CHANGE IN SIGN

Suppressor effects sometimes caused a change in the sign of a relationship with the dependent variable. A study of crime patterns provides an example of change in sign from the zero-order correlation of a predictor and dependent variable to the partial correlation and regression coefficients of the predictor with the dependent variable. This study examined information on socioeccnomic and demographic characteristics of the population in conjunction with crime rates for census tracts in Kansas City, Missouri. The model used in analyzing the data we will discuss below was as follows:



where X is a socioeconomic or demographic indicator,
R is the rate for robbery, and
L is the rate for larceny

The model is intended to determine the extent to which knowledge of robbery rates and social information can help predict tracts in which larceny rates will be high or lcw. As shown in Table 1, four different social indicators (two measures of family structure, one measure of occupational status, and one measure of racial composition) were tried as the X variable in alternate stepwise regressional runs. The question being asked here was whether the X variable affects larceny rates, either directly, or indirectly through its effect on robbery rates.

As shown in Table 1, the zero-order relationships between the social variables and larceny are not significant with the exception of the family-structure variable, Percent of Adult Females Separated. However, the relationships for the three

non-significant predictors become significant as measured by both the partial correlation coefficients and the regression coefficients when controls are imposed for robbery rates. Morever, the relationship between Percent of Adult Females Separated remains significant when robbery is controlled, but the sign changes from positive to negative. Robbery acts as a suppressor between all of the X variables and larceny.

We generally have been taught to regard the sign of a correlation of a regression coefficient as the measure of the direction of association between two variables. The sign of the relationship between Percent of Adult Females Separated changes from significantly positive in the bivariate relationships to significantly negative when robbery rates are controlled in the multivariate situation. Decomposing the correlation coefficient in accordance with the causal model again shows how this happens:

$$r_{LX} = B_{LX.R} + B_{RX}B_{LR.X}$$

$$.20286 = -.12462 + .38630\ (.84773)$$

$$= -.12462 + .32748$$

$$= .20286$$

The reader may well ask what has been gained by understanding how suppression works in this example of data on relationships between larceny, robbery, and social information for urban census tracts. The answer is that knowledge of suppressor effects can help improve our understanding and predictions of phenomena in the real world. In the present example, an analyst interested in identifying census tracts likely to have high larceny rates based on the data at hand would note the high correlation between larceny and robbery (.79959) and conclude that 64 percent of the variance in the former variable was predictable from the latter. Looking at the non-significant zero-order correlation between larceny and Percent Female-head Families, the analyst might conclude that the latter variable could offer no additional help in improving the prediction. This conclusion

would be a mistake, of course, because Percent Female-head Families has a significant relationship with larceny after controlling for robberty, and it improves the R^2 in predicting larceny by six percent ($-.24286^2$).

Related to this point is the substantive issue embodied in the finding that larceny has a negative relationship with the social indicators when controlling for robbery. Why, for example, should larceny rates be lower in census tracts with a high percent of Female-head families than in tracts with a low percent when controls are imposed for robbery or other possible indicators of social disorganization? The answer is not intuitively obvious, and we do not at this time have a good answer to that question, but the analyst who is at least aware of the question has a lead useful in working to identify and eliminate the conditions that result in high larceny rates. An examination of the zero-order correlations without an awareness and understanding of the suppressor relationships in the data set would not have provided this lead.

CONCLUSION

We have reviewed examples of suppressor variables in multiple regression research in which 1) one predictor accounts for more of the variance in the dependent variable when another predictor is included; and 2) the last variable to enter an equation accounts for more variance than the next-to-last variable. The second situation is a sub-category of the first but it is useful to look at the two separately to understand how suppression operates in multiple regression.

In addition, special attention was given to: 1) a case in which analysis of variance helps to clarify suppression; 2) a case in which suppressors are part of a larger set of predictor variables; and 3) a case in which the relationship between a predictor and the dependent variable changes in size when another predictor is controlled.

Our purpose in reviewing and illustrating these situations was to clarify how and why suppression operates and what to look for in identifying and understanding it. Many researchers who use stepwise multiple regression programs limit all or most of their examination of statistical output to zero-order correlations which may provide a misleading portrayal of the relationships between variables. In particular, researchers sometimes find that variables with a low zero-order correlation with the dependent add more to the explained variance than the square of that correlation. A researcher unfamiliar with suppression may attribute this pattern to "blow-up" in the statistical program due to small N's or multicollinearity among predictors. This may be the case but the analyst also should study the partial correlation and regression coefficients to determine whether the pattern may be a real one attributable to suppression.

If so, the researcher also may find that suppression helps to unravel and clarify substantive relationships among variables. For example, our analysis

using the data set on number of physicians related to hospital beds and expenditures suggested that facilities expenditures do help to attract physicians but only if one takes into account these variables' relationships with hospital beds. Understanding of suppressor effects can provide leads which the analyst will find worth pursuing in attempting to clarify how and why the predictors account for variance in the dependent variable, and the circumstances in which these effects can be expected to operate most forcefully.

Table 1

Interrelationships of Four Alternate Social Indicators (X Variables), Robbery Rates, (R), and Larceny Rates (L) + for Census Tracts in Kansas City, Missouri (N = 161)

X Variable	r_{LX}	$r_{LX.R}$	$B_{LX.R}$	$B_{LR.X}$	B_{RX}
Percent Negro	-.07272	-.21974**	-.13232**	.80934**	.07364
Percent Adult Female-Separated	.20286*	-.19140*	-.12462*	.84773**	.38630**
Percent of Workers Who are Laborers	.11730	-.18578*	-.11624*	.83222**	.28063*
Percent Female-head Families	-.06008	-.24286**	-.14668**	.81518**	.10623

+B_{LR} = .79959

*P < .05

** P < .01

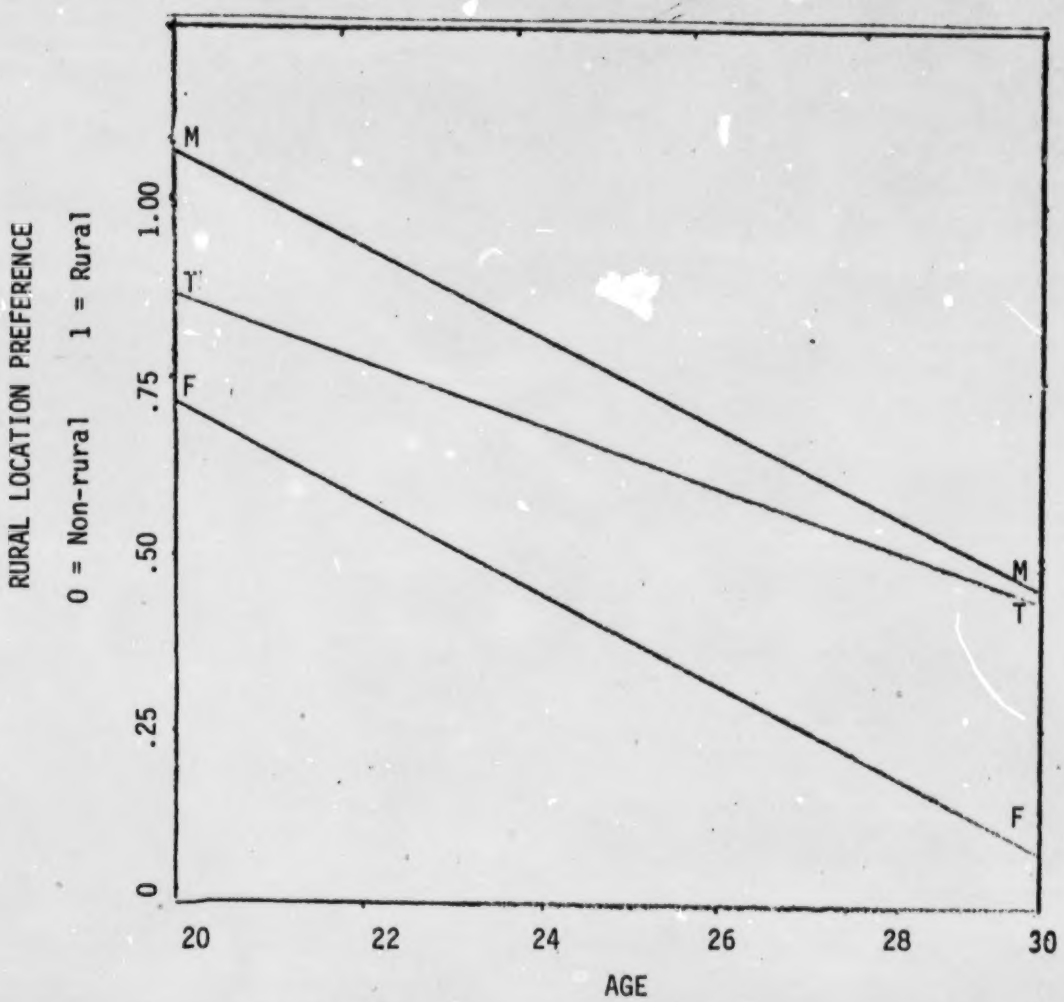


Figure 1. Regression Lines of the Predicted Scores on Rural Location Preference and Age. M=Male, F=Female, T=Total group.